SCIENTIFIC REPORTS
nature research

OPEN

Link clustering explains non-central and contextually essential genes in protein interaction networks

Received: 1 May 2019
Accepted: 1 August 2019
Published online: 12 August 2019

Inhae Kim[1], Heetak Lee[1], Kwanghwan Lee[1], Seong Kyu Han[1], Donghyo Kim[1] & Sanguk Kim[1,2]

Recent studies have shown that many essential genes (EGs) change their essentiality across various contexts. Finding contextual EGs in pathogenic conditions may facilitate the identification of therapeutic targets. We propose link clustering as an indicator of contextual EGs that are non-central in protein-protein interaction (PPI) networks. In various human and yeast PPI networks, we found that 29–47% of EGs were better characterized by link clustering than by centrality. Importantly, non-central EGs were prone to change their essentiality across different human cell lines and between species. Compared with central EGs and non-EGs, non-central EGs had intermediate levels of expression and evolutionary conservation. In addition, non-central EGs exhibited a significant impact on communities at lower hierarchical levels, suggesting that link clustering is associated with contextual essentiality, as it depicts locally important nodes in network structures.

A gene is essential for cell viability when its loss-of-function is lethal to the cell. Gene essentiality is contextual, as a gene can change its essentiality across different species, growth media, and chemical treatments[1–6]. There is growing evidence that some essential genes (EGs) are more contextual than others. For instance, some yeast EGs were more evolvable (i.e., able to become non-essential) than others under laboratory conditions[7]. In addition, genome-scale loss-of-function screening across cancer cell lines revealed that a subset of genes were less often essential than others[8–10]. The identification of such contextual EGs holds great potential for the development of therapies that selectively suppress pathogenic cells by targeting genes that are essential in the pathogenic cells but not in healthy cells[11].

Several reports have shown that invariable EGs are likely to be central nodes in protein-protein interaction (PPI) networks, whereas contextual EGs are less central. For instance, human genes that were broadly essential across cell lines showed greater degree than other genes in PPI networks[10,12]. In addition, yeast genes that remained essential during laboratory evolution showed greater degree than those that became non-essential[7]. An intuitive explanation for those observations is that central EGs are indispensable for their implication on global network integrity, which is imperative regardless of varying contexts that alter the pertinence of local regions (Fig. 1, red area). By contrast, non-central EGs can be indispensable in one context but dispensable in another context (Fig. 1, blue area). One implication of that hypothesis is that the characterization of non-central EGs would lead to the identification of contextual EGs.

Network clustering is a property that can characterize non-central EGs distinctly from central EGs. Previous studies showed that EGs tend to be directly connected to one another[13,14] or enriched within the same functional modules and protein complexes[15–20], thus making up essential modules. It has not been tested whether such clustered EGs are contextual. In addition, the question of whether a clustering measure could separate non-central EGs from central ones also needs to be explored. The nodes that make up essential modules tend to have greater degree than the nodes in other modules[20]. Furthermore, essential modules themselves have greater numbers of connections than non-essential modules in module-level networks[21]. A few studies have looked specifically at EGs with high link clustering[22,23], but they tested the distinction between non-central EGs and central EGs only weakly by comparing 100 genes with highly ranked topology measures.

We hypothesize that highly clustered links, rather than merely a clustered network structure, contribute to gene essentiality, because such links represent functional dependency between nodes. It was previously proposed

[1]Department of Life Sciences, Pohang University of Science and Technology, Pohang, 37673, Korea. [2]School of Interdisciplinary Bioscience and Bioengineering, Pohang University of Science and Technology, Pohang, 37673, Korea. Correspondence and requests for materials should be addressed to S.K. (email: sukim@postech.ac.kr)

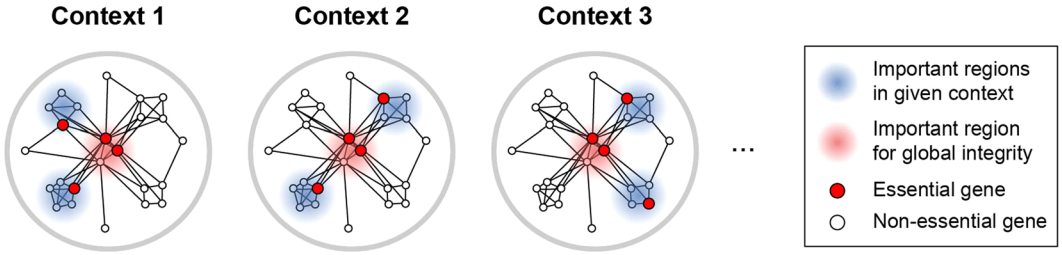

Figure 1. Hypothesis about the relationship between network structure and contextual essentiality. Non-central EGs might be essential in certain contexts, whereas central EGs would be essential for their role in global integration regardless of context.

that links with stronger functional dependency have a greater impact on network robustness than links with weaker functional dependency, as the failure of one node with strong functional dependency will likely result in the failure of the whole neighborhood[24,25]. At the molecular level, obligate interactions among proteins are one example of functional dependency: a protein is unstable on its own, so it has to be bound to its partner to sustain its stability[26]. It has been shown that various link clustering measures can estimate functional dependency between nodes[27–29].

We aimed to characterize the relationship between contextual EGs and non-central EGs. We systematically compared various clustering measures for their ability to characterize non-central EGs and investigated their association with contextual EGs. We found that link clustering is an accurate indicator of node essentiality independent of centrality, enabling us to correctly classify a substantial number of non-central EGs. EGs with clustered links were likely to change their essentiality across human cell lines and between species and, furthermore, showed levels gene expression and evolutionary conservation that were between those of central EGs and non-EGs. Moreover, the non-central EGs had profound impacts on communities at low-level hierarchy, supporting our hypothesis that network clustering is relevant to contextual essentiality because it characterizes locally pertinent nodes in the network.

Results

Centrality and link clustering characterize distinct facets of gene essentiality. Our goal was to find a clustering measure that is capable of characterizing EGs that are distinct from central EGs, which we classify as non-central EGs (see Fig. S1 for a flowchart). We investigated a node clustering measure (node clustering coefficient, *C*) and three link clustering measures (the product of two end nodes' *C* values, *CXC*; the link clustering coefficient, *LCC*; and the edge clustering coefficient, *ECC*). Because gene essentiality is a property of nodes rather than of links, for each node we aggregated the link clustering measures with the node's neighbors in PPIs by taking the average (μCXC, μLCC, and μECC) and the sum (ΣCXC, ΣLCC, and ΣECC). We compared those clustering measures to four centrality measures (degree, *DC*; betweenness, *BC*; closeness, *CC*; and eigenvector, *EC*) in eight different PPI networks in yeast and human. The details of the topology measures and PPI networks are further described in the *Supplementary Information* (*SI*).

We found that several of the link clustering measures are as capable of characterizing gene essentiality as the centrality measures are (Fig. 2A). For each topology measure, we divided proteins into rank-ordered bins and calculated Pearson's correlation coefficient (*R*) between the given measure and the fraction of EGs (f_E). Five link clustering measures, including μCXC, μLCC, ΣCXC, ΣLCC, and ΣECC, showed significant correlations with f_E in most of the PPI networks, whereas *C* and μECC often failed to exhibit significant correlations (Fig. 2A; gray $^*P < 0.05$; black $^*P < 0.001$). In addition, we confirmed that the centrality measures (*DC*, *BC*, *CC*, and *EC*) also exhibited significant correlations with f_E in various PPI networks, which is consistent with many previous reports.

The average link clustering measures (μCXC and μLCC) were more distinct from the centrality measures than the sum measures (ΣCXC, ΣLCC and ΣECC), although they were all correlated with f_E. We found that μCXC and μLCC mostly exhibited no correlation with the centrality measures (Fig. 2B; blue versus red), whereas ΣCXC, ΣLCC, and ΣECC were often strongly correlated with the centrality measures (Fig. 2B; green versus red). To explore the relationship among topology measures more comprehensively, we conducted a principal component analysis (PCA) of EGs. In the yeast consolidated network, the average link clustering measures (Fig. 2C; 2–4, blue arrows) were roughly orthogonal to the centrality measures (Fig. 2C; 8–11, red arrows), whereas the sum link clustering measures (Fig. 2C; 5–7, green arrows) were prominently oriented in a similar direction with the centrality measures. We observed similar results in other PPI networks (Fig. S2). Therefore, the sum measures were not suitable for our goal to find non-central EGs, because they seemed to depict central EGs. All of the correlations between topology measures and gene essentiality, and their statistical significance, are shown in Table S1.

Because μCXC and μLCC were correlated with f_E (Fig. 2A) and not with the centrality measures (Fig. 2B,C), we examined them further for distinction from the centrality measures. We conducted a pruning analysis in which we removed nodes in decreasing order of a given topology measure and monitored the resultant change in excess clustering (ΔC), the difference between the observed *C* and the average *C* of random networks subjected to degree sequence-preserved randomization. The pruning analysis provided a comparison of topology measures in terms of their implication on network clustering.

We found that μCXC had greater implication on network clustering than μLCC and, more importantly, was more distinct from the centrality measures than μLCC. In the human consolidated network, for instance, the

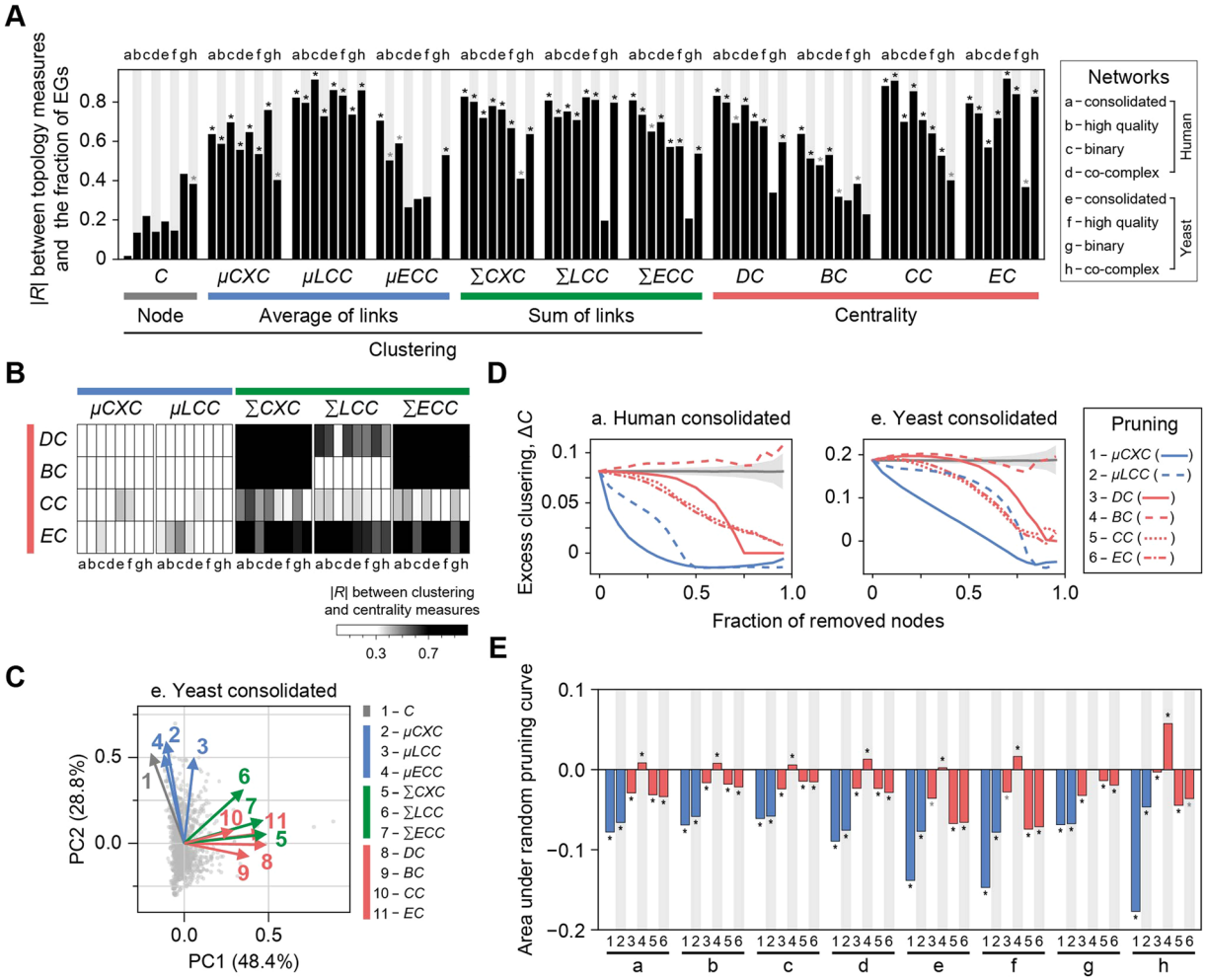

Figure 2. Relationships between topology measures and gene essentiality in PPI networks. (**A**) The absolute value of the Pearson correlation coefficient (|*R*|) between the fraction of EGs and topology measures. Proteins were sorted by a given topology measure and divided into bins each containing 2% of the population. (**B**) |*R*| between centrality measures and clustering measures. (**C**) Principle component (PC) analysis of topology measures on EGs in the yeast consolidated network. The variance explained by each component is given in parentheses. (**D**) Pruning analysis. The change in excess clustering (ΔC) was monitored while proteins were progressively removed in decreasing order of a given topology measure, or randomly (gray line; area, 3σ). ΔC is the difference between the observed C and the mean C of randomized networks. (**E**) Summary of pruning analyses in different networks. The decrease of ΔC was quantified by the area under the random pruning curve.

pruning curve of *μCXC* exhibited a slightly faster decrease of ΔC than that of *μLCC* (Fig. 2D, left; line 1 versus line 2), indicating that *μCXC* had a somewhat stronger impact on network clustering. In that case, both link clustering measures were distinguishable from the centrality measures, as the centrality measures showed much slower decreases of ΔC (Fig. 2D, left; lines 3–6). By contrast, in the yeast consolidated network, the pruning curve of *μLCC* (Fig. 2D, right; line 2) overlapped those of two centrality measures, *CC* and *EC* (Fig. 2D, right; lines 5 and 6, respectively), indicating that those three topology measures had similar impacts on network clustering. We quantified the decrease of ΔC by measuring the area over the curve of each parameter and under that of random pruning (Fig. 2D, gray line), in which proteins were removed in a random order. We observed that *μCXC* was distinct from the centrality measures in all the PPI networks, whereas *μLCC* was similar to the centrality measures in the yeast consolidated, high-quality, and co-complex networks (Fig. 2E; see Fig. S3 for all the pruning curves).

Since *μCXC* was the most distinct from the centrality measures and capable of characterizing gene essentiality, we used it to classify non-central EGs. For the sake of simplicity, we refer to *μCXC* as *w* throughout the rest of the manuscript, as it represents the link weights. We selected *DC* as a counterpart to classify central EGs and refer to it as *k*.

Link clustering characterizes a distinct subset of non-central EGs. Given that gene essentiality can be characterized by two uncorrelated properties, *k* and *w*, we expect EGs to fall into two distinct subsets: those better characterized by *k* (*k*-dependent) and those better characterized by *w* (*w*-dependent). Using logistic regression, we calculated the probabilities of being essential based on *k*, $P_E(k)$, and based on *w*, $P_E(w)$. We then classified

EGs as *k*-dependent if $P_E(k) > P_E(w)$ or as *w*-dependent if $P_E(k) < P_E(w)$ (Fig. S4). Considering the cases where EGs are explained by neither *k* nor *w*, we discarded EGs under cutoffs k_c and w_c, which maximized Matthew's correlation coefficient (MCC) by regarding only genes with $k \geq k_c$ or $w \geq w_c$ as predicted EGs (Fig. S5).

We found that a sizable number of EGs were *w*-dependent. In the human consolidated network, 36.0% of EGs were *w*-dependent ($n = 2{,}186$; Fig. 3A, left; blue circles), which is comparable to the proportion of *k*-dependent EGs (40.9%, $n = 2{,}483$; Fig. 3A, left; red circles). Those two subsets of EGs were very distinctive in the network structure (Fig. 3B, left). As expected, the *w*-dependent EGs showed greater *w* than the *k*-dependent EGs (Fig. 3B, right; $P = 4.4 \times 10^{-68}$, Mann-Whitney U [MWU] test) and the non-EGs ($P = 0$), and they had intermediate *k* compared with the *k*-dependent EGs ($P = 0$) and the non-EGs ($P = 5.5 \times 10^{-108}$). In the eight different PPI networks, 29–47% of EGs were *w*-dependent (Fig. 3C). All of the *k*-dependent and *w*-dependent EGs in different PPI networks are shown in Tables S3, S4 for yeast and human, respectively.

We next examined whether *k*-dependent and *w*-dependent EGs are distinct with respect to not only network structure but also biological function. We defined *k*-functions and *w*-functions as gene ontology (GO) terms enriched with *k*-dependent and *w*-dependent EGs, respectively, in three or four PPI networks. Because different GO terms could be similar to each other, we constructed functional networks of GO terms connected by shared genes and investigated function-clusters in which GO terms were densely connected (see the *Methods*).

We found that *k*-dependent and *w*-dependent EGs were associated with distinct biological functions. In the yeast functional network composed of Cellular Components (CC) terms, many function-clusters were biased toward either *k*-functions or *w*-functions (Fig. 3D; see Fig. S6 for all the functional networks). For instance, one function-cluster was composed of four similar *k*-functions ("septin complex", "mating projection base", "septin filament array", and "cellular bud neck septin ring") that were enriched with *k*-dependent EGs from all four yeast PPI networks (Fig. 3D; box C1). Another cluster possessed three related *w*-functions ("ribonuclease MRP complex", "telomerase holoenzyme complex", and "nucleolar ribonuclease P complex") that were enriched with *w*-dependent EGs from all four yeast PPI networks (box C2). In addition, links in the functional network were observed more frequently between pairs of *k*-functions (Fig. 3E, upper panel; $n = 359$, $z = 2.12$) and between pairs of *w*-functions (Fig. 3E, lower panel; $n = 292$, $z = 3.34$) compared with those observed in 10,000 random sets with shuffled *k*-function and *w*-function tags. By contrast, links between *k*-functions and *w*-functions were observed at a frequency similar to the random expectation (Fig. 3E, middle panel; $n = 401$, $z = -0.38$). That result was robust over all functional networks composed of three GO categories in yeast and human (Fig. S7).

We made a comprehensive summary of the biological functions associated with *k*-dependent and *w*-dependent EGs (Fig. 3F for yeast; Fig. S8 for human). We selected a GO term with the median size, as determined by the number of genes assigned to the term, as a representative for each function-cluster. Among 45 function-clusters in the yeast functional networks, only four ("ATP-dependent RNA helicase activity", "RNA polymerase I activity", "DNA-directed RNA polymerase II, core complex", and "core TFIIH complex") were biased toward both *k*-functions and *w*-functions ($-\log[P] \geq 2$, hypergeometric test). By contrast, 15 and 14 function-clusters were biased toward either *k*-functions or *w*-functions, respectively. Function-clusters biased toward *k*-functions often represented cytokinesis (e.g., "septin complex", "establishment or maintenance of actin cytoskeleton polarity", and "mitotic cytokinesis"), whereas those biased toward *w*-functions corresponded to RNA degradation (e.g., "exosome [RNase complex]", "ribonuclease MRP complex", and "nuclear polyadenylation-dependent mRNA catabolic process"). Taken together, those results demonstrate that link clustering characterizes a unique subset of EGs with distinct biological functions. All *k*-functions and *w*-functions and their clusters in yeast and human are shown in Tables S5 and S6, respectively.

_w_-dependent EGs are more contextual than _k_-dependent EGs. There is growing evidence that gene essentiality is often contextual, meaning that a gene may change its essentiality across cell lines and species. Given that central genes tend to be evolutionarily conserved and expressed broadly across cell lines, we expect that *w*-dependent EGs might be prone to change their essentiality, as they are less central than *k*-dependent EGs.

As expected, we found that the essentiality of *w*-dependent EGs was more cell-line-specific than that of *k*-dependent EGs (Fig. 4A). Using a publicly available dataset of genetic vulnerability screens in 436 cancer cell lines, we measured the broadness of essentiality for each gene as the number of cell lines in which the given gene exhibited a fitness-effect ≤ -0.3. We then divided the EGs into five mostly even bins and observed the distributions of *k*-dependent and *w*-dependent EGs. In the human consolidated network, the *w*-dependent EGs tended to be essential in a relatively small number of cell lines, whereas the *k*-dependent EGs were essential in a greater number of cell lines (Fig. 4A; $P = 4.0 \times 10^{-26}$, χ^2 test). Supporting that observation, we also found that *w*-dependent EGs exhibited an intermediate level of expression between those of *k*-dependent EGs and non-EGs (Fig. 4B,C). With the expression dataset matched to the genetic vulnerability screens, *w*-dependent EGs tended to be expressed less broadly among the cell lines than *k*-dependent EGs (Fig. 4B, $P = 1.0 \times 10^{-70}$) and more broadly than non-EGs ($P = 5.0 \times 10^{-130}$). In addition, the average expression level of *w*-dependent EGs was lower than that of *k*-dependent EGs (Fig. 4C, $P = 4.5 \times 10^{-26}$) and higher than that of non-EGs ($P = 1.0 \times 10^{-85}$). We observed similar results in other PPI networks with different cutoffs (Figs S9–11), except in the binary network.

We also found that the essentiality of *w*-dependent EGs was more frequently changed between human and mouse than that of *k*-dependent EGs. In the human consolidated network, mouse orthologs of *w*-dependent EGs were more frequently identified as non-essential (Fig. 4D; fraction = 50.8%) than those of *k*-dependent EGs (35.0%, $P = 4.0 \times 10^{-16}$, Fisher's exact test), indicating that the essentiality of *w*-dependent genes was less conserved than that of *k*-dependent EGs. We also found that *w*-dependent EGs exhibited an intermediate level of molecular conservation compared with *k*-dependent EGs and non-EGs (Fig. 4E,F). We estimated protein ages on the basis of a reconstructed history of protein families and found that *w*-dependent EGs were younger than *k*-dependent EGs (Fig. 4E; $P = 3.7 \times 10^{-40}$, χ^2 test) and older than non-EGs ($P = 2.9 \times 10^{-72}$). In addition, we observed that the evolutionary rate (dN/dS, ratio of synonymous to non-synonymous nucleotide substitutions)

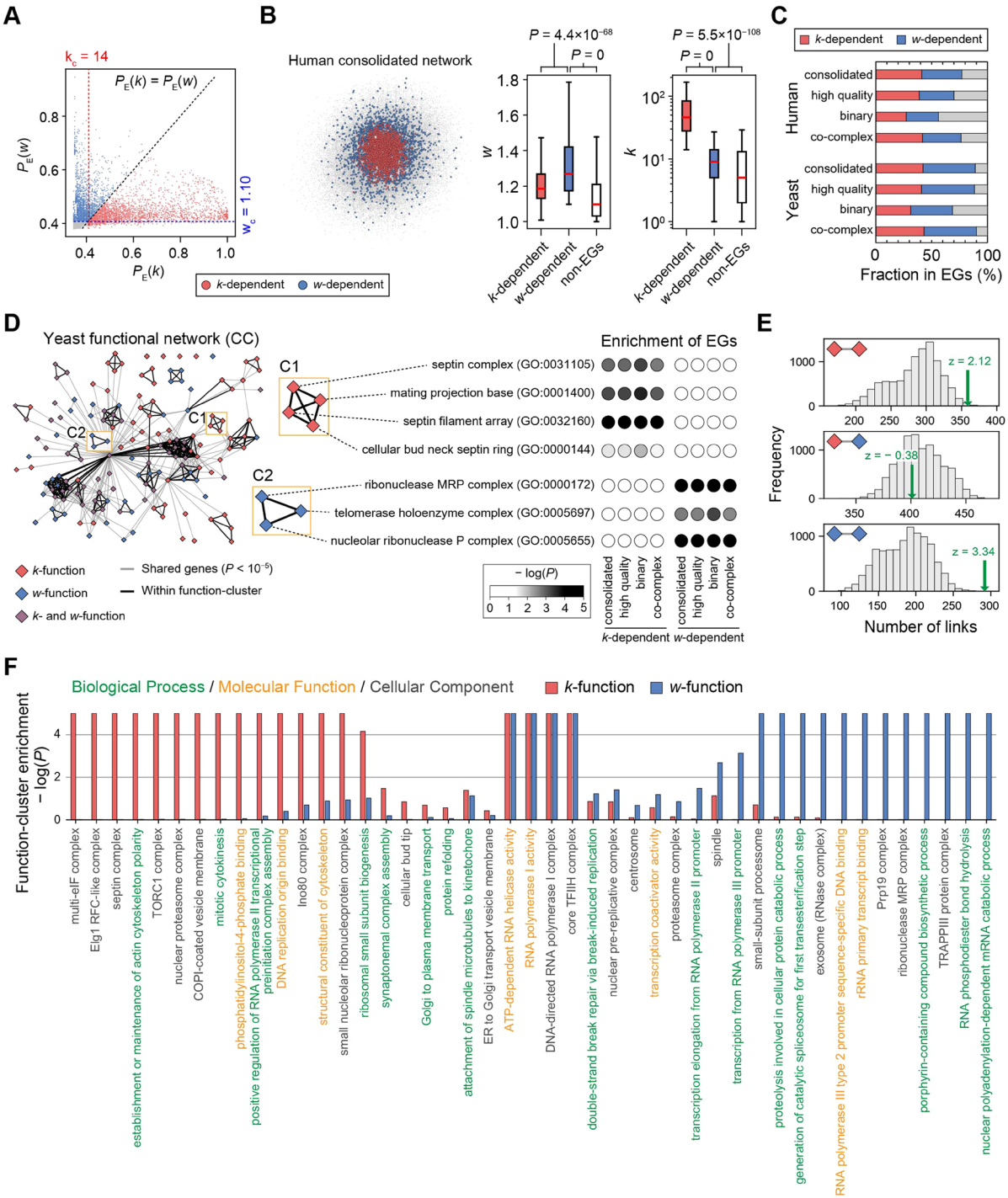

Figure 3. Classification of *k*-dependent and *w*-dependent EGs and their functional differences. (**A**) Classification of *k*-dependent and *w*-dependent EGs based on the probability of being essential (P_E) inferred from logistic regression with *k* and *w*. Cutoffs (k_c and w_c) were determined to maximize MCC for each topology measure, and proteins under the cutoffs remained unclassified. (**B**) (left) *k*-dependent and *w*-dependent EGs in the network. (right) Topological differences among classified EGs and non-EGs. (**C**) Fraction of *k*-dependent and *w*-dependent EGs in different PPI networks. (**D**) The functional network of yeast "cellular component" (CC) terms connected by shared genes. Terms were identified as *k*-functions or *w*-functions when they were enriched with *k*-dependent or *w*-dependent EGs, respectively, in three or four PPI networks. Function-clusters were determined by the MCODE algorithm. (**E**) Number of links between *k*-functions and *w*-functions in the functional network of yeast CC terms; (upper) between two *k*-functions; (middle) between a *k*-function and a *w*-function; (lower) between two *w*-functions. Green arrows indicate the observed number in the real network, and gray bars show the number distribution in 10,000 random sets with shuffled *k*-functions and *w*-functions. (**F**) Representative GO terms of function-clusters from yeast functional networks and the bias of clusters toward *k*-functions and *w*-functions. *P*-values $< 10^{-5}$ were set to 10^{-5}.

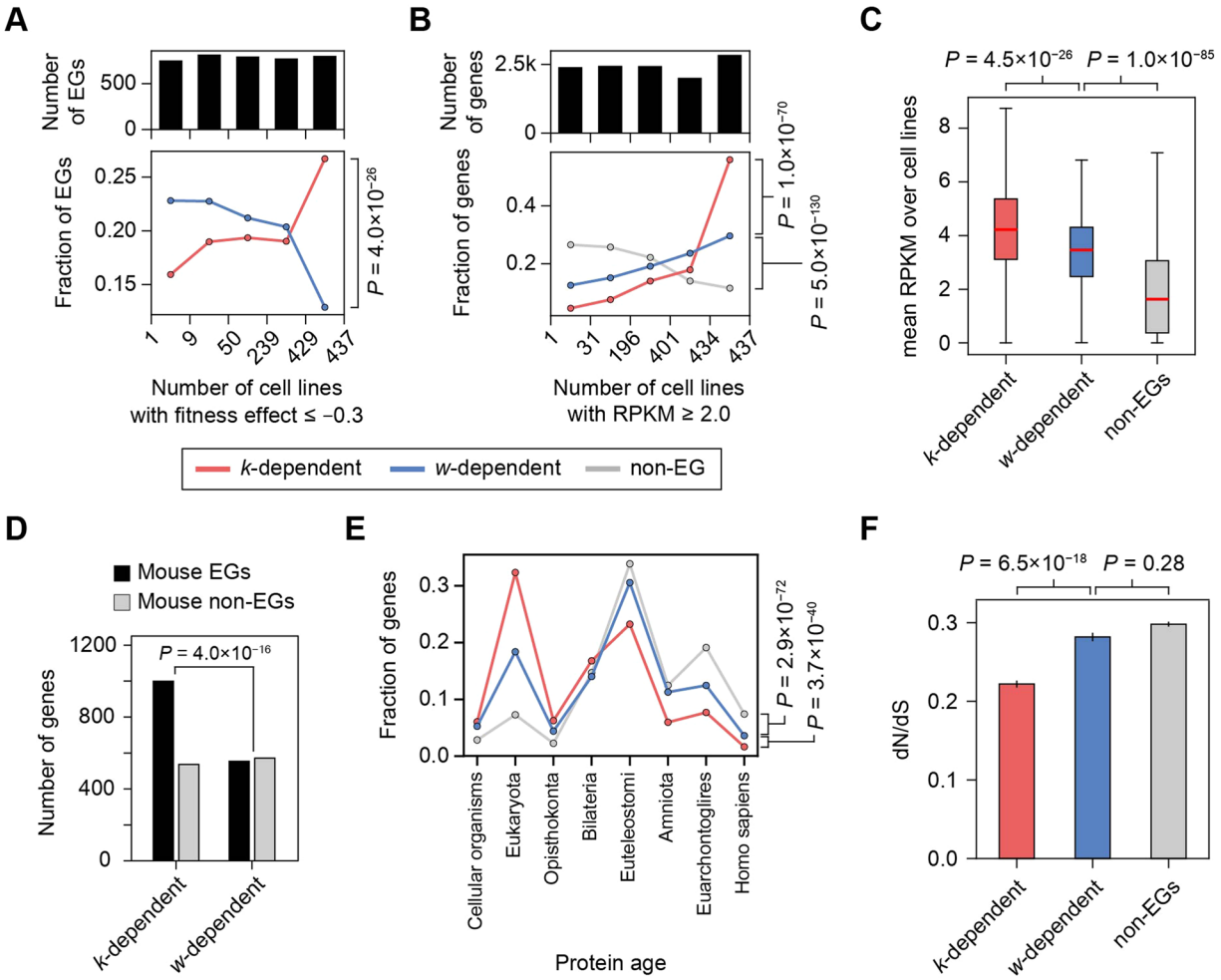

Figure 4. Contextual essentiality of human *k*-dependent and *w*-dependent EGs. (**A**) Contextual gene essentiality across human cell lines for *k*-dependent and *w*-dependent EGs. Bins were divided to have similar populations of EGs. (**B**) Contextual gene expression across human cell lines. Bins were divided to have similar populations of all genes. (**C**) Expression levels of genes across human cell lines. (**D**) Essentiality change of *k*-dependent and *w*-dependent EGs between mouse and human. (**E**) Phyletic age and (**F**) evolutionary rate (dN/dS) of *k*-dependent and *w*-dependent EGs and non-EGs.

of *w*-dependent EGs was greater than that of *k*-dependent EGs (Fig. 4F; $P = 6.5 \times 10^{-18}$, MWU test) and less than that of non-EGs, although the latter difference was not statistically significant. Similar results were observed in other PPI networks with different cutoffs (Figs S12, S13), except in the binary network.

Taken together, the results strongly suggest that *w*-dependent EGs are more contextual than *k*-dependent EGs. The investigated molecular properties, such as gene expression and evolutionary conservation, only characterized *w*-dependent EGs as analogous to non-EGs (Fig. 4B,C,E,F), whereas the link clustering showed *w*-dependent EGs as being further apart from non-EGs than from *k*-dependent EGs (Fig. 3B).

***w*-dependent EGs significantly impact communities at low levels of hierarchy.** Because many previous studies have already suggested network clustering as a property pertinent to gene essentiality, we examined *k*-dependent and *w*-dependent EGs more precisely regarding the clustered network structure around them. Specifically, we found that *k*-dependent EGs are well-clustered in a generic sense, whereas *w*-dependent EGs are specifically relevant to network communities at low levels of hierarchy.

We found that *k*-dependent EGs were more clustered than *w*-dependent EGs (Fig. 5A–C). There were more *k*-functions than *w*-functions (Fig. 5A), indicating that *k*-dependent EGs are more likely to be clustered into the same functions than *w*-dependent EGs. Because the difference in the numbers of *k*-dependent and *w*-dependent EGs might affect the observed enrichment, we randomly removed the same number of EGs from each category and monitored the decrease of enriched GO terms. In the yeast consolidated network, for instance, the removal of *k*-dependent EGs lead to a greater decrease in the number of enriched GO terms than the removal of *w*-dependent EGs in all three GO categories (Fig. 5B). To further examine the clustered network structure around EGs, we searched *n*-cliques, which are fully connected subgraphs with *n* nodes, and investigated their bias toward *k*-dependent or *w*-dependent EGs. We found that cliques frequently included more *k*-dependent EGs than *w*-dependent EGs (Fig. 5C; 3-cliques, fraction = 43.4% versus 13.0%, $z = -485.0$; 4-cliques, fraction = 42.3%

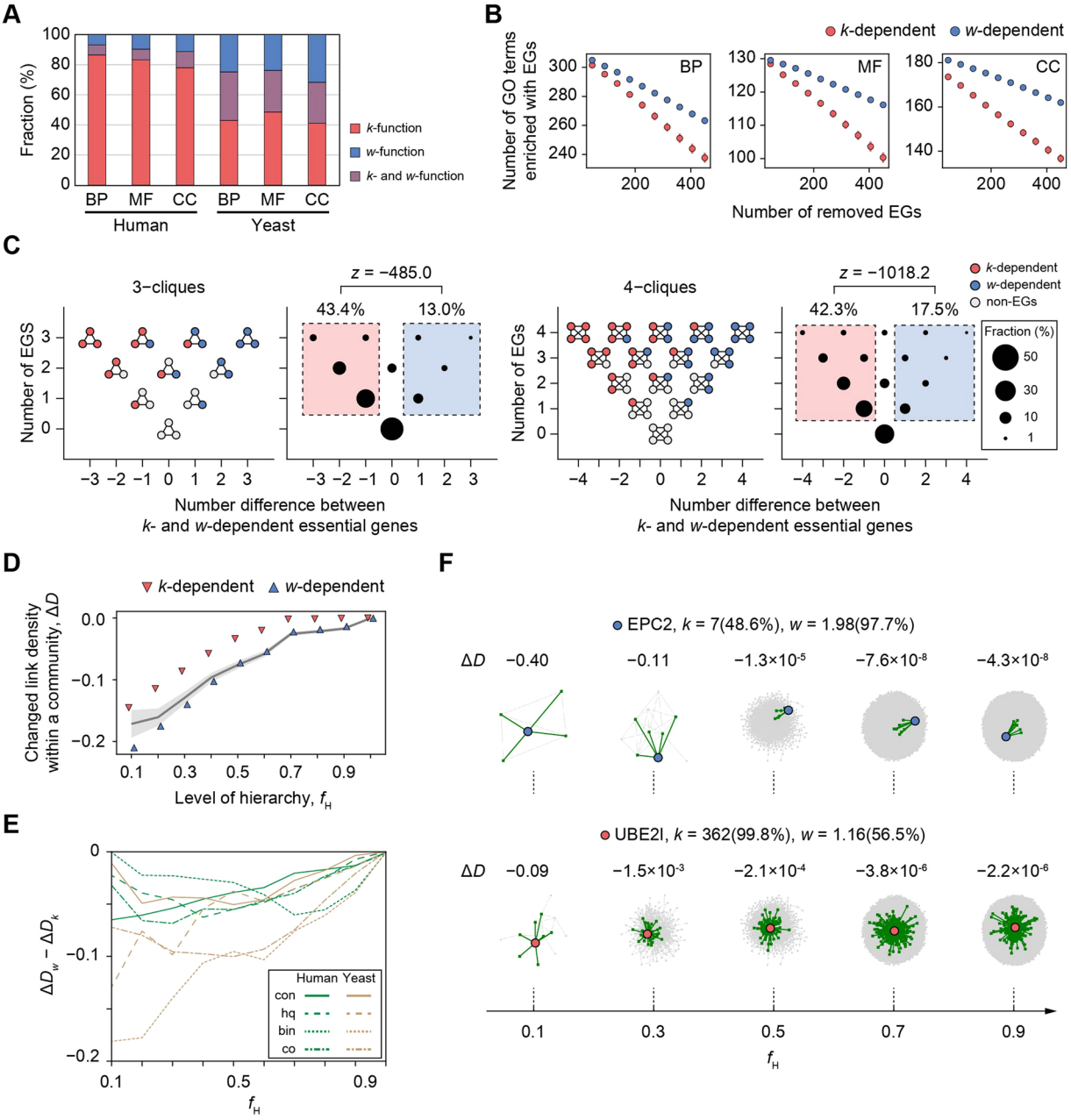

Figure 5. Clustering of *k*-dependent and *w*-dependent EGs. (**A**) Relative frequency of GO terms enriched with *k*-dependent and *w*-dependent EGs. (**B**) Decrease of enriched GO terms due to removal of the same number of *k*-dependent or *w*-dependent EGs. (**C**) *n*-cliques and their biases toward *k*-dependent and *w*-dependent EGs. The normal distribution of *z* values was approximated by binomial trials with success probability given by the fraction of cliques biased toward *k*-dependent EGs. (**D**) Change in the link density of a community upon removal of a single node (ΔD) at different hierarchical levels (f_H). Hierarchical community structure was searched using the *Walktrap* algorithm. The gray line indicates the average ΔD for non-EGs (area, standard error). (**E**) Difference in ΔD between *k*-dependent and *w*-dependent EGs in different PPI networks. (**F**) Examples illustrating *EPC2*, a *w*-dependent EG, and *UBE2I*, a *k*-dependent EG, for their impact on communities. The protein of interest (blue circles, *EPC2*; red circles, *UBE2I*) and its interactions (green lines) with its first neighbors (green circles) are shown in the community. Numbers in parentheses indicate the rank percentile for *k* and *w*.

versus 17.5%, $z = -1018.2$). Similar results were observed in other PPI networks, except in the binary networks (Figs S14–S15). Those results indicate that *k*-dependent EGs are densely clustered into the same biological functions, possibly because of their greater number of links.

The observed clustering of *k*-dependent EGs raises the question of how the link clustering measure, *w*, separates a subset of non-central EGs. Given their contextual essentiality, we hypothesized that *w*-dependent EGs would significantly impact small communities at low levels of hierarchy, because a system's dependency on a

small and local community would be more context-specific than that on a large and global community. To test that hypothesis, we investigated the impact of the removal of a node by monitoring changes of link density, ΔD, in communities at various hierarchical levels, f_H, defined by the fraction of prior merges in the agglomerative clustering (see the *Methods*).

We found that the impact on link density within a community was greater for the removal of *w*-dependent EGs than for that of *k*-dependent EGs, and the difference was significant at lower hierarchical levels. In the human consolidated network with $f_H = 0.1$, we observed a greater decrease in link density for *w*-dependent EGs (Fig. 5D, $\Delta D_w = -0.210$) than for *k*-dependent EGs ($\Delta D_k = -0.145$) upon removal of a single node, suggesting that *w*-dependent EGs have a greater impact on community structure. By contrast, the difference in ΔD between *k*-dependent and *w*-dependent EGs became extremely small at the highest hierarchical level ($\Delta D_w - \Delta D_k = -0.00052$, $f_H = 1.0$), suggesting that the effect of a single node removal is unlikely to be distinguishable at the level of the global network. Additionally, to confirm that ΔD is relevant to gene essentiality, we looked at changes in link density upon the removal of a single node for non-EGs (ΔD_n). At lower hierarchical levels ($f_H \leq 0.4$), $\Delta D_w - \Delta D_n < 0$, indicating that *w*-dependent EGs had a greater impact on local community structure than non-EGs. Similar results were observed in other PPI networks; $\Delta D_w - \Delta D_k < 0$ (Fig. 5E) and $\Delta D_w - \Delta D_n < 0$ (Fig. S17) at lower hierarchical levels with low f_H values.

An example of the impact of single node deletions on community structure with varying hierarchy is shown in Fig. 5F. At $f_H = 0.1$, the deletion of *EPC2*, a *w*-dependent EG with few and clustered links ($k = 7$ [rank percentile = 48.6%], $w = 1.98$ [rank percentile = 97.7%]) had a large impact on the community structure ($\Delta D = -0.40$), causing the removal of four of nine total links. By contrast, the deletion of *UBE2I*, a *k*-dependent EG with many unclustered links ($k = 362$ [rank percentile = 99.8%], $w = 1.16$ [rank percentile = 56.5%]) had a smaller impact on the community structure ($\Delta D = -0.09$), although more links ($n = 7$) were removed. At higher hierarchical levels ($f_H \geq 0.5$), both *EPC2* and *UBE2I* became members of the same large communities, so the impact of their deletion was much smaller ($\Delta D_{EPC2} = -1.3 \times 10^{-5}$ and $\Delta D_{UBE2I} = -2.1 \times 10^{-4}$, at $f_H = 0.5$) than at lower levels of hierarchy. Taken together, those results indicate that both *k*-dependent and *w*-dependent EGs could be considered "clustered" in some sense, whereas the link clustering discretely characterizes EGs that are crucial for communities at low levels of hierarchy.

Discussion

We demonstrated that a link clustering measure, *w*, is capable of characterizing non-central and contextual EGs. For the understanding of contextual gene essentiality, the biological significance of link clustering measures remains a matter of scientific exploration. Our results strongly suggest that functional dependency between nodes, rather than network clustering per se, is crucial for depicting contextual EGs. We observed that *w*-dependent EGs have distinct implications on communities at low levels of hierarchy (Fig. 5D–F), in which strong functional relevance among member nodes is expected. Recent reports showed that a gene's essentiality across varying contexts is largely dependent on its neighbors with strong functional relevance[30–32]. Moreover, links conveying strong functional dependency may have a significant impact on network robustness, as the failure of one node will likely cascade over them[24,25]. Taken together, our results suggest that the link clustering measure *w* estimates functional dependency between two nodes and portrays genes that are functionally pivotal to their neighbors in non-central regions of cellular systems.

Many previous studies suggested relevance between gene essentiality and network clustering, so one might reasonably ask whether non-central EGs were characterized in those studies[13–20]. It is worth noting that "clustering" is a general concept of network structure, and a clustering measure may or may not distinguish non-central nodes from central ones. In our dataset, we observed that some clustering measures, other than *w*, were correlated with centrality measures (Fig. 2A–C), and that central EGs were in some sense "clustered" with respect to functional modules and network cliques (Fig. 5A–D). Therefore, in working toward the goal to separate non-central EGs from central EGs, one needs to carefully assess different topology measures, as each measure characterizes a distinct facet of the network structure.

The limitation of our work is that the link clustering measure *w* is incapable of estimating a gene's essentiality for a given context, despite its ability to characterize the tendency for a gene to be contextual across contexts. Precise estimation of gene essentiality for a given cell line has potential for the development of therapeutic targets that specifically eliminate pathogenic cells without causing excessive damage to normal cells[9,10]. In addition, recent genome-scale fitness screens enabled the identification of molecular biomarkers for contextual essentiality, providing insights into the molecular mechanisms underlying the vulnerability of pathogenic cells[33,34]. We anticipate that the further classification of EGs will provide a useful indication of varying gene essentiality in different contexts.

It has been argued that disease genes are devoid of essentiality and network centrality, because the impairment of an EG would likely cause the death of the organism rather than manifest disease phenotypes[35]. Associations between EGs and diseases might be more prevalent than expected[36,37], however, because many genetic perturbations that occur naturally may not be as severe as the complete loss-of-function induced in gene essentiality assays in the laboratory. For instance, human genes with mouse-essential orthologs were likely to be associated with the manifestation of severe and life-threatening diseases[38]. In addition, non-coding RNAs (ncRNAs), which often target and regulate the expression of hub proteins in PPI networks[39], exhibited profound relevance to various biological pathways and diseases[40,41], suggesting that they have rather tolerable implications on EGs. With respect to microRNA-mediated diseases, we observed that *w*-dependent EGs were associated with more dissimilar diseases than *k*-dependent EGs, although the numbers of associated diseases were not significantly different (Fig. S18). Therefore, with the growing resources for investigating ncRNAs and their relevance to diseases[42–45], we expect that explorations of the association between stratified EGs and ncRNAs will provide useful insights into the molecular etiology of diseases.

One might ask whether the topology measures used here are robust to incompleteness of the networks. To answer that question, we investigated the correlations between k, w, and f_E in 100 random networks with 50% of the links removed. We found that both k and w were robust to the random changes of links (Fig. S19). With the removal of links, the correlations of f_E with k and w were only slightly decreased, and the difference was insignificant in most networks. In addition, the correlation between k and w remained close to 0, indicating that k and w would characterize distinct subsets of EGs. Therefore, we expect that our results will remain robust in more complete networks in the future.

Although our goal was to categorize different topology measures, one might also integrate various measures for the classification of EGs from non-EGs. Indeed, we found that the combination of centrality and clustering measures improved the power to predict EGs (see the *SI*; Fig. S20), although we simply used the rank of the topology measures as the predictive parameters. That strongly suggests that the application of more complicated statistical models or machine learning algorithms will further improve the prediction of EGs. In particular, recent studies have demonstrated that deep learning is a powerful approach to model complex genotype-phenotype associations[46,47], providing insights into gene-disease associations[48] and polypharmacy side-effects[49]. We believe that the incorporation of stratified topology measures with deep learning will improve the prediction of gene essentiality.

Regarding a gene as the unit of evolution, a gene might be selfish[50] and establish a large number of clustered links with strong functional dependency ("strong links"), rendering itself indispensable for many cellular functions and thus ensuring its persistence in the population. In fact, many previous studies suggested a similar interpretation: that, rather than being crucial for global integration, central EGs simply have a greater chance to be involved in essential functions[13,17,20]. We observed that such selfishness is constrained, however; central EGs seemed to have weaker links than non-central EGs (Fig. 3B). From a systems perspective, a gene's selfishness would not always be tolerable, as it comes at a fitness costs to the population. Assuming a system with such selfish genes of promiscuous functional relevance, a random failure may not be properly insulated, and the system would not be resilient to the frequent random errors in non-central nodes. This systems perspective asserts that strong links are constrained from connecting central nodes to other nodes. Indeed, strong links were found to be likely confined within local regions in various real networks including PPI networks[51,52], genetic interaction networks[53], brain connectomes[54], and social networks[28]. That suggests that gene essentiality evolves in a tradeoff between a gene's importance and its implication on system robustness, and one needs to synthesize gene-centric and system-centric perspectives for a comprehensive understanding of gene essentiality.

Methods

Relationships between gene essentiality and topology measures. The "consolidated" PPI networks were downloaded from the web interface to the Interaction Reference Index repository (iRefWeb)[55] on June 7, 2017. The "binary" and "co-complex" networks were downloaded from the high-quality interactomes (HINT) database[56] on June 27, 2017. The "high-quality" networks were created by combining the binary and co-complex networks. Gene essentiality information was downloaded from the online gene essentiality (OGEE) database[57] on June 9, 2017. Any essentiality annotation with the "TextMining" data type was removed.

To explore a parameter's ability to characterize gene essentiality, we calculated the Pearson correlation coefficient (R) between the fraction of EGs (f_E) and the average of a given parameter along with the rank-ordered groups. Proteins were sorted by increasing order of the parameter of interest and added into a single bin until the bin contained at least 2% of the total population. This applied to all relationships of f_E with centrality measures (DC, BC, CC, and EC) and clustering measures (C, μCXC, μLCC, μECC, ΣCXC, ΣLCC, and ΣECC). See Table S7 in the *SI* for the definitions of topology measures.

For the PCA of EGs, we used the decomposition.PCA() object in the Python "scikit-learn" package. To scale the features, we also used the transform() function of the preprocessing.StandardScaler() object in the same package.

Monitoring global and local connectivity upon pruning. Pruning analysis was performed in a manner similar to that previously reported[29]. Proteins were progressively removed from a given network at 5% of the total protein population in decreasing order of k and w while corresponding changes in ΔC with varying f (the fraction of removed proteins) were monitored. To calculate the ΔC of individual nodes, we constructed 100 random networks by degree sequence-preserved randomization[29,58] and subtracted the mean node clustering coefficients of random sets from the observed node clustering coefficient.

To summarize the results of the pruning analyses, we measured the area between a given curve and the random curve. Because the decrease in ΔC was our interest, we specifically measured the area under the random curve. Therefore, if the given curve was over the random curve, the measured area became negative. We linearly interpolated the curves and calculated the trapezoidal area over $f = [0, 0.95]$.

Classification of *k*-dependent and *w*-dependent EGs. A gene could only be classified one of two ways: essential or non-essential. We assigned values of 1 to EGs and 0 to non-EGs. The probability that a given gene is essential was then calculated using logistic regression analysis according to a leave-one-out scheme, with k and w as dependent variables, resulting in $P_E(k)$ and $P_E(w)$, respectively. We performed the logistic regression analysis using the Python "scikit-learn" package. In addition, k_c (w_c) was determined to maximize MCC regarding all nodes with $k \geq k_c$ ($w \geq w_c$) as predictive positives.

Functional association between *k*-dependent and *w*-dependent EGs. To construct functional networks, we defined GO terms as k-functions if they were enriched with k-dependent genes in at least three PPI networks ($P < 0.05$, hypergeometric test); we also defined w-functions accordingly. GO terms were discarded

when the number of genes annotated to them was less than three. Note that a GO term could be enriched with both k-dependent and w-dependent EGs, as the two types of enrichment were tested independently. A link was established between two GO terms if there was significant gene overlap ($P < 10^{-5}$) between them. We used the MCODE application[59] to identify clusters in the functional networks. For each cluster, we selected the median-sized GO term as the cluster's representative function, where size is the number of genes annotated to the function. We constructed a total of six functional networks for three GO categories (BP, MF, and CC) and two eukaryotic species (yeast and human). Annotations were downloaded from the GO database[60,61]; the submission date of the human data used in the study was September 26, 2017, and that of yeast data was September 13, 2017.

Impact of node removal on community structure. For each PPI network, we constructed a hierarchical organization based on the *Walktrap* algorithm[62], using the Python package "python-igraph". We chose the algorithm for its concept underlying the similarity between nodes. The algorithm relies on random walks to measure the similarity between two nodes by comparing their probability of random visits on other nodes: if two nodes are in the same community, then random walks starting from each node will visit all the other nodes in the same way. This process is somewhat reminiscent of the failure cascade shown in recent works, in which a single node failure was propagated and resulted in system-wide catastrophe[63]. After the similarity between nodes is established, the clustering process is agglomerative. In the earlier steps, nodes with greater similarity are put together into a community. Therefore, we took the fraction of prior merge steps, f_H, as an indication of hierarchy; the smaller the f_H, the lower the hierarchical level. By increasing f_H by 0.1 in a step-wise manner, we collected communities at different levels in the hierarchical organization. Communities comprising less than three members were discarded. The change in link density in community s upon the deletion of node i was calculated as follows: $\Delta D_{s,i} = (l_s - l_{s,\Delta i})/(n_s \times (n_s - 1)/2)$, where l_s denotes the number of links within s (i.e., two end nodes are both members of s), $l_{s,\Delta i}$ represents the number of links within s after removing node i, and n_s is the number of members in s. Therefore, ΔD measures the proportion of links removed upon a node deletion, indicating the extent of functional dependency within a community relying on the deleted node.

Essential genes across contexts. For human cancer cell lines, we used a CRISPR screen dataset including 436 cell lines (gene_effect.csv, 18Q2) from DepMap database[64]. The expression level of genes was downloaded from the CCLE database[65] with the matching version.

The essentiality of mouse genes was downloaded from the OGEE database[57], similarly to that of human genes. Orthologs between human and mouse were identified from the Inparanoid database[66] (version 8.0). Gene ages were downloaded from the ProteinHistorian database[67]; specifically, we used protein families predicted from the OrthoMCL and PANTHER databases and reconstructed ancestral history by asymmetric Wagner parsimony. We used the pre-calculated set of dN/dS for yeast[68] and human[69], for which evolutionary rates were computed with several species and the average taken.

Diseases associated with miRNAs. The relationships between genes and disease were constructed by connecting gene-miRNA and miRNA-disease associations. For gene-miRNA associations, we used miRTarBase,[70] discarding pairs with only "weak" evidence. For miRNA-disease associations, we used two different databases, HMDD[71] and MDGHI[45]. Because MDGHI is a predictive approach, we applied an arbitrary cutoff and discarded all pairs with score smaller than 0.01.

Data Availability

All data generated or analyzed during this study are provided in this published article and its *Supplementary Information* files, and at sbi.postech.ac.kr/w/WEG.

References

1. Nichols, R. J. *et al*. Phenotypic Landscape of a Bacterial Cell. *Cell* **144**, 143–156 (2011).
2. Steinmetz, L. M. *et al*. Systematic screen for human disease genes in yeast. *Nat. Genet.* **31**, 400–404 (2002).
3. Chen, S., Zhang, Y. E. & Long, M. New Genes in Drosophila Quickly Become Essential. *Science* **330**, 1682–1685 (2010).
4. Hillenmeyer, M. E. *et al*. The Chemical Genomic Portrait of Yeast: Uncovering a Phenotype for All Genes. *Science* **320**, 362–365 (2008).
5. Mnaimneh, S. *et al*. Exploration of Essential Gene Functions via Titratable Promoter Alleles. *Cell* **118**, 31–44 (2004).
6. Lee, A. Y. *et al*. Mapping the Cellular Response to Small Molecules Using Chemogenomic Fitness Signatures. *Science* **344**, 208–211 (2014).
7. Liu, G. *et al*. Gene Essentiality Is a Quantitative Property Linked to Cellular Evolvability. *Cell* **163**, 1388–1399 (2015).
8. Blomen, V. A. *et al*. Gene essentiality and synthetic lethality in haploid human cells. *Science* **350**, 1092–1096 (2015).
9. Wang, T. *et al*. Identification and characterization of essential genes in the human genome. *Science* **350**, 1096–1101 (2015).
10. Hart, T. *et al*. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163**, 1515–1526 (2015).
11. Rancati, G., Moffat, J., Typas, A. & Pavelka, N. Emerging and evolving concepts in gene essentiality. *Nat. Rev. Genet.* **19**, 34–49 (2018).
12. Bertomeu, T. *et al*. A High-Resolution Genome-Wide CRISPR/Cas9 Viability Screen Reveals Structural Features and Contextual Diversity of the Human Cell-Essential Proteome. *Mol. Cell. Biol.* **38**, 1–24 (2017).
13. He, X. & Zhang, J. Why Do Hubs Tend to Be Essential in Protein Networks? *PLoS Genet.* **2**, e88 (2006).
14. Lu, C. *et al*. Why do essential proteins tend to be clustered in the yeast interactome network? *Mol. Biosyst.* **6**, 871 (2010).
15. Hart, G. T., Lee, I. & Marcotte, E. R. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics* **8**, 236 (2007).
16. Luo, J. & Qi, Y. Identification of Essential Proteins Based on a New Combination of Local Interaction Density and Protein Complexes. *PLoS One* **10**, e0131418 (2015).
17. Ryan, C. J., Krogan, N. J., Cunningham, P. & Cagney, G. All or Nothing: Protein Complexes Flip Essentiality between Distantly Related Eukaryotes. *Genome Biol. Evol.* **5**, 1049–1059 (2013).
18. Semple, J. I., Vavouri, T. & Lehner, B. A simple principle concerning the robustness of protein complex activity to changes in gene expression. *BMC Syst. Biol.* **2**, 1 (2008).

19. Wang, H. *et al*. A Complex-based Reconstruction of the Saccharomyces cerevisiae Interactome. *Mol. Cell. Proteomics* **8**, 1361–1381 (2009).
20. Zotenko, E., Mestre, J., O'Leary, D. P. & Przytycka, T. M. Why Do Hubs in the Yeast Protein Interaction Network Tend To Be Essential: Reexamining the Connection between the Network Topology and Essentiality. *PLoS Comput. Biol.* **4**, e1000140 (2008).
21. Song, J. & Singh, M. From Hub Proteins to Hub Modules: The Relationship Between Essentiality and Centrality in the Yeast Interactome at Different Scales of Organization. *PLoS Comput. Biol.* **9**, e1002910 (2013).
22. Wang, J., Li, M., Wang, H. & Pan, Y. Identification of Essential Proteins Based on Edge Clustering Coefficient. IEEE/ACM Trans. *Comput. Biol. Bioinforma.* **9**, 1070–1080 (2012).
23. Li, M., Zhang, H., Wang, J. & Pan, Y. A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. *BMC Syst. Biol.* **6**, 15 (2012).
24. Parshani, R., Buldyrev, S. V. & Havlin, S. Critical effect of dependency groups on the function of networks. *Proc. Natl. Acad. Sci.* **108**, 1007–1010 (2011).
25. Bashan, A., Parshani, R. & Havlin, S. Percolation in networks composed of connectivity and dependency links. *Phys. Rev. E* **83**, 051127 (2011).
26. Nooren, I. M. A. Diversity of protein-protein interactions. *EMBO J.* **22**, 3486–3492 (2003).
27. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. & Parisi, D. Defining and identifying communities in networks. *Proc. Natl. Acad. Sci.* **101**, 2658–2663 (2004).
28. Onnela, J.-P. *et al*. Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci.* **104**, 7332–7336 (2007).
29. Pajevic, S. & Plenz, D. The organization of strong links in complex networks. *Nat. Phys.* **8**, 429–436 (2012).
30. Pan, J. *et al*. Interrogation of Mammalian Protein Complex Structure, Function, and Membership Using Genome-Scale Fitness Screens. *Cell Syst.* **6**, 555–568.e7 (2018).
31. Boyle, E. A., Pritchard, J. K. & Greenleaf, W. J. High-resolution mapping of cancer cell networks using co-functional interactions. *Mol. Syst. Biol.* **14**, e8594 (2018).
32. Hart, T., Koh, C. & Moffat, J. Coessentiality And Cofunctionality: A Network Approach To Learning Genetic Vulnerabilities From Cancer Cell Line Fitness Screens. *bioRxiv* **134346**, https://doi.org/10.1101/134346 (2017).
33. Tsherniak, A. *et al*. Defining a Cancer Dependency Map. *Cell* **170**, 564–576.e16 (2017).
34. Behan, F. M. *et al*. Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. *Nature* **568**, 511–516 (2019).
35. Goh, K.-I. *et al*. The human disease network. *Proc. Natl. Acad. Sci.* **104**, 8685–8690 (2007).
36. Spataro, N., Rodríguez, J. A., Navarro, A. & Bosch, E. Properties of human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology. *Hum. Mol. Genet.* **26**, ddw405 (2017).
37. Georgi, B., Voight, B. F. & Bućan, M. From Mouse to Human: Evolutionary Genomics Analysis of Human Orthologs of Essential Genes. *PLoS Genet.* **9**, e1003484 (2013).
38. Han, S. K., Kim, I., Hwang, J. & Kim, S. Network Modules of the Cross-Species Genotype-Phenotype Map Reflect the Clinical Severity of Human Diseases. *PLoS One* **10**, e0136300 (2015).
39. Hsu, C.-W., Juan, H.-F. & Huang, H.-C. Characterization of microRNA-regulated protein-protein interaction network. *Proteomics* **8**, 1975–1979 (2008).
40. Chen, X., Yan, C. C., Zhang, X. & You, Z.-H. Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* **18**, bbw060 (2016).
41. Chen, X., Xie, D., Zhao, Q. & You, Z.-H. MicroRNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* **20**, 515–539 (2019).
42. Chen, X. & Yan, G.-Y. Novel human lncRNA–disease association inference based on lncRNA expression profiles. *Bioinformatics* **29**, 2617–2624 (2013).
43. Chen, X. & Huang, L. LRSSLMDA: Laplacian Regularized Sparse Subspace Learning for MiRNA-Disease Association prediction. *PLOS Comput. Biol.* **13**, e1005912 (2017).
44. Chen, X., Wang, L., Qu, J., Guan, N.-N. & Li, J.-Q. Predicting miRNA–disease association based on inductive matrix completion. *Bioinformatics* **34**, 4256–4265 (2018).
45. Chen, X., Yin, J., Qu, J. & Huang, L. MDHGI: Matrix Decomposition and Heterogeneous Graph Inference for miRNA-disease association prediction. *PLOS Comput. Biol.* **14**, e1006418 (2018).
46. Li, Y. *et al*. Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. *Methods*, https://doi.org/10.1016/j.ymeth.2019.04.008 (2019).
47. Ma, J. *et al*. Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods* **15**, 290–298 (2018).
48. Li, Y., Kuwahara, H., Yang, P., Song, L. & Gao, X. PGCN: Disease gene prioritization by disease and gene embedding through graph convolutional neural networks. *bioRxiv* **532226**, https://doi.org/10.1101/532226 (2019).
49. Zitnik, M., Agrawal, M. & Leskovec, J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* **34**, i457–i466 (2018).
50. Dawkins, R. *The selfish gene*. (Oxford: Oxford University Press 1976).
51. Hein, M. Y. *et al*. A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. *Cell* **163**, 712–723 (2015).
52. Kim, I., Lee, H., Han, S. K. & Kim, S. Linear Motif-Mediated Interactions Have Contributed to the Evolution of Modularity in Complex Protein Interaction Networks. *PLoS Comput. Biol.* **10**, e1003881 (2014).
53. Costanzo, M. *et al*. A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, aaf1420–aaf1420 (2016).
54. Gallos, L. K., Makse, H. A. & Sigman, M. A small world of weak ties provides optimal global integration of self-similar modules in functional brain networks. *Proc. Natl. Acad. Sci.* **109**, 2825–2830 (2012).
55. Turner, B. *et al*. iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database* **2010**, baq023–baq023 (2010).
56. Das, J. & Yu, H. HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst. Biol.* **6**, 92 (2012).
57. Chen, W.-H., Lu, G., Chen, X., Zhao, X.-M. & Bork, P. OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. *Nucleic Acids Res.* **45**, D940–D944 (2017).
58. Maslov, S. Specificity and Stability in Topology of Protein Networks. *Science* **296**, 910–913 (2002).
59. Bader, G. D. & Hogue, C. W. V. Analyzing yeast protein–protein interaction data obtained from different sources. *Nat. Biotechnol.* **20**, 991–997 (2002).
60. Ashburner, M. *et al*. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
61. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45**, D331–D338 (2017).
62. Pons, P. & Latapy, M. Computing Communities in Large Networks Using Random Walks. *J. Graph Algorithms Appl.* **10**, 191–218 (2006).
63. Buldyrev, S. V., Parshani, R., Paul, G., Stanley, H. E. & Havlin, S. Catastrophic cascade of failures in interdependent networks. *Nature* **464**, 1025–1028 (2010).

64. Meyers, R. M. *et al*. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49**, 1779–1784 (2017).
65. Stransky, N. *et al*. Pharmacogenomic agreement between two cancer cell line data sets. *Nature* **528**, 84–87 (2015).
66. Sonnhammer, E. L. L. & Östlund, G. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.* **43**, D234–D239 (2015).
67. Capra, J. A., Williams, A. G. & Pollard, K. S. ProteinHistorian: Tools for the Comparative Analysis of Eukaryote Protein Origin. *PLoS Comput. Biol.* **8**, e1002567 (2012).
68. Chakraborty, S. & Ghosh, T. C. Evolutionary Rate Heterogeneity of Core and Attachment Proteins in Yeast Protein Complexes. *Genome Biol. Evol.* **5**, 1366–1375 (2013).
69. Chakraborty, S., Panda, A. & Ghosh, T. C. Exploring the evolutionary rate differences between human disease and non-disease genes. *Genomics* **108**, 18–24 (2016).
70. Chou, C.-H. *et al*. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* **46**, D296–D302 (2018).
71. Lu, M. *et al*. An Analysis of Human MicroRNA and Disease Associations. *PLoS One* **3**, e3420 (2008).

Acknowledgements

This work was supported in part by Korean National Research Foundation grants (2018R1A2B6002657 and 2017M3C9A60472625) and a Korea Institute of Marine Science & Technology grant (D11510215H480000140).

Author Contributions

I.K. and S.K. designed the study; I.K., H.L., S.H. and K.L. collected and analyzed the data; I.K., D.K., H.L., built the webpage; I.K. and S.K. wrote the paper.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-48273-3.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019